VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

M.tech(Software Engineering)

Fall Semester (2020-21)

TOPIC : - DIABETIC PREDICTION USING ENSEMBLE ALGORITHM

COURSE NAME : - DESIGN PATTERNS

COURSE CODE : - SWE2019

SLOT : - C1

FACULTY : SENTHIL KUMAR .M

REVIEW-3

Submitted By:

R.HARISH-17MIS0476

P.PRADEEP-17MIS0410

Introduction:-

Ensemble modeling is a powerful way to improve the performance of your model. It usually pays off to apply ensemble learning over and above various models you might be building. Time and again, people have used ensemble models in competitions like Kaggle and benefited from it. Ensemble learning is a broad topic and is only confined by your own imagination. For the purpose of this article, I will cover the basic concepts and ideas of ensemble modeling. This should be enough for you to start building ensembles at your own end. As usual, we have tried to keep things as simple as possible. Let's quickly start with an example to understand the basics of Ensemble learning. This example will bring out, how we use ensemble model every day without realizing that we are using ensemble modeling.
Ensemble is the art of combining diverse set of learners (individual models) together to improvise on the stability and predictive power of the model. In the above example, the way we combine all the predictions together will be termed as Ensemble Learning. In this article, we will talk about a few ensemble techniques widely used in the industry. Before we get into techniques, let's first understand how do we actually get different set of learners. Models can be different from each other Diabetes mellitus (DM) or simply diabetes, is a group of metabolic diseases

Existing system:

Develop an optimal Bayesian neural network algorithm for the detection of hypoglycemia episodes in T1DM children using physiological parameters such as heart rate, corrected QT interval and skin impedance. Hypoglycemiaorlow blood glucose is a common and serious side effect of insulin therapy in patients with diabetes[1]

Diabetic autonomic, neuropathy (DAN) is common complication of diabetes mellitus. (DM), associated with development of angiopathies and increased mortality. The spectral power of low frequency (LF), high frequency(HF) and total frequency(TF) components of HRV was significantly reduced in DM patients as compared to the control.[2]

The method proposed here uses genetic programming (GP) and a variation of genetic programming called GP with comparative partner selection (CPS) for diabetes detection. In first stage we use genetic programming to produce an individual from training data, that converts the available features to a single feature such that it has different values for healthy and patient(diabetes) data. In the next stage we use test data for testing of that individual.[3]

The proposed datamining technique is classification. For the detection of diabetes the approach used data mining algorithms. Based on the complexity of the attributes and based on the characteristics of dataset, selection of this method Is done. Classification is the major essential decision making techniques in real world situation.[4]

There port suggested from different health organization shows the alarming condition due to diabetes worldwide. This paper analyses and provides the related implications on the previous research work. It highlighted the advantages and the missing areas of the previous study. It provides the direction informing the new prediction frame work design.[5]

Architecture of neural network is developed. This back propagation neural network structure has one hidden layer with 10 neurons in the hidden layer, 8 input nodes and 1 output node as the problem as this structure used for binary classification. The output could be 0 or 1, which shows 0 as normal patient and 1 as diabetic. The results as diabetic and non-diabetic can be displayed on screen.[6]

Identification of MA at an early stage is the best solution for the early discovery of DR. Here, classification of digital retinal fundus images is performed by employing One rule and back propagation neural networks for two classes namely diabetic or non-diabetic. The retinal fundus image is partitioned in to four equal parts which makes it a better approach as compared to the methods reported in the literature due to the availability of a large number of features.[7]

Applied the stacked ensemble (or super learning) data mining method to predict the short-term vs. long-term length of hospital stay of hospitalized patients with diabetes. The median LOS retrieved from the processed diabetic patients' dataset was used as the threshold dividing short-term vs. long-term LOS.[8]

Present early detection system of DM by using four models of artificial neural networks. Performance of artificial neural network models for early detection of DM using Confusion Matrix. The results concluded that the use of a combination of Backpropagation method with PSO optimization and ALR can solve the problems in the outliers in the data well with the best accuracy compared to other methods.[9]

Explores the possibility of detection of diabetes in the initial stage. For this different methods are explored from the literature and the past research works. This provides a detail analysis and overview of different methodologies of data mining. This also includes comparison, method analysis and gap identification.[10]

DM detection system named Computer-assisted Non-invasive Diabetes Mellitus Detection System (CNDMDS) is designed and developed to help medical professionals quickly and easily to detect DM in real time.CNDMDS has two parts: (a) a non-invasive device used to capture facial images and (b) a software installed in the computer connected with this device detecting DM and showing the results in real time.[11]

Predicting diabetes by applying data mining technique. The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in dataset and generate clear and understandable description of patterns.[12]

:Huge number of individuals are influenced by Diabetes Mellitus (DM) which is hard to cure because of its endless nature and hereditary connection. The uncontrolled diabetes may prompt heart related issues.[13]

various machine learning algorithms are used to predict diabetes and specific attributes. The performances of the algorithms are compared in terms of accuracy, voting based ensemble techniques is applied for the normalized pima diabetes data for which a highest accuracy is achieved.[14]

Diabetes Prediction as graphical format. Here it is considered that predicted value which closer to 1 and above .5 is considered as 1 (Positive) whereas, closer to 0 and below .5 is considered as 0 (Negative)Prediction capacity of ANN based model in can predict the possibility of developing diabetes in the community of Pima Indians.[15]

The disease predictions have been explored using various methods of data mining. The use of medical data set on the prediction of diabetic mellitus has been analysed. This paper performs a detailed survey on disease prediction using data mining approaches based on diabetic data set. The presence of disease has been identified using the appearance of various symptoms.[16]

Diabetes mellitus is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both. Early diagnosing of diabetic causing heart, kidney and eye complications is difficult and challenging. Data mining techniques are applied on clinical data attributes of diabetics to predict the risk factors. The aim of the paper is to develop a fuzzy classification model to predict heart and kidney complications using diabetic clinical data.[17]

Diabetic retinopathy (DR) is the most common cause of newly diagnosed blindness every year. DR detection is solely based on existing patient records. Now a day’s medical data growing tremendously and we need to process that data for detection. But it is time consuming hence data mining techniques helps to get rid from this issue. We use neural network (NN) and naïve bayes for classification. According to comparison results NN gives better accuracy than naïve bayes and time and memory required for NN is less as compared to naïve bayes.[18]

In this study, the retinal image is taken from a fundus camera of both healthy and diabetic retina. Image pre-processing techniques, morphological operations are used to detect the statistical features and the histogram-based feature is extracted by using Discrete Wavelet Transform (DWT) which is the novel contribution of the proposed algorithm. These features are classified by any machine learning approach (K-Nearest Neighbors, Support Vector Machine and Artificial Neural

Network) to predict DR accurately and efficiently following a cross-validation approach.[19]

Segmentation of vascular structures of retina for implementation of Clinical diabetic retinopathy decision making systems is presented in this paper. As retinal vascular structure is with thin blood vessels, prediction accuracy is highly dependent upon the segmentation and preprocessing schemes.Binarization algorithms are used to achieve the segmentation of vascular structures of the retina.[20]

The main problem in this type of dieses is its prediction. It is found that if diabetes mellitus is detected at early stages then it can be cured. So early detection of diabetes mellitus is important. There are different techniques with the help of which early detection of diabetes mellitus is possible. In this paper combination of three different methods used for early detection of diabetes mellitus are given. These three methods are fuzzy system, neural network, case based reasoning. By using combination of all these approaches, it is found that detection of diabetes mellitus at early stages is possible.[21]

Diabetic retinopathy (DR) diagnosis methods in the literature are usually criticized as being limit in diagnosing DR-related features or being lack of interpretability. To deal with these issues, this paper investigates the feasibility of diagnosing both DR severity levels and the presence of DR-related features in a two-step procedure. Specifically, this paper first analyzes the quality of annotations in DR grading by measuring inter-grader variability.[22]

Diabetes is known as a metabolic disease. Type 1 diabetes is an anti-immune disease whereby the body's immune system kills off its own insulin producing beta cells in the pancreas. Type 2 diabetes is an advanced state of health in which the body becomes opposed to the usual impacts of insulin and/or progressively loses the capacity to produce adequate amount insulin in the pancreas, and it finally may not be able to produce any insulin. Type 2 diabetes is the most common form of the disease with complications including heart, vision, and foot conditions.[23]

Diabetes, also known as chronic illness, is a group of metabolic diseases due to a high level of sugar in the blood over a long period. The risk factor and severity of diabetes can be reduced significantly if the precise early prediction is possible. The robust and accurate prediction of diabetes is highly challenging due to the limited number of labeled data and also the presence of outliers (or missing values) in the diabetes datasets.[24]

Diabetic Retinopathy (DR) is an eye disease due to diabetes, which is the most ordinary cause of blindness among adults of working age in Malaysia. To date, DR is still screened manually by ophthalmologist using fundus images due to insufficiently reliable existing automated DR detection systems. This paper proposed an algorithm that consists of DR detection method with the aim to improve the accuracy of the existing systems.[25]

Diabetes mellitus is one of the most common chronic diseases. The number of cases of diabetes in the world is likely to increase more than two fold in the next 30 years; from 115 million in 2000 to 284 million in 2030. In type I diabetes, the disease is caused by the failure of the pancreas to produce a sufficient amount of insulin which leads to an uncontrolled increase in blood glucose unless the patient administers insulin, typically by subcutaneous injection.[26]

This paper presents an improved diabetic retinopathy detection scheme by extracting accurate area and ate number of microaneurysm from color fundus images. Regular screening of eye is crucial for detection and dealing with diabetic retinopathy. Diabetic retinopathy (DR) is an eye disease which occurs due to damage of retina as a result of long illness of diabetic mellitus.[27]

Diabetic Retinopathy (DR), a major complication of diabetes and the leading cause of new cases of blindness among adults, can be cured by the early and precise detection of the disease. An important aspect of DR is the micro- vascular changes that cause detectable changes in the appearance of retinal blood vessels. In this paper, we propose a new blood-vessel detection technique in retinal images, based on the regional recursive hierarchical decomposition using Quadtrees and post-filtration of edges[28]

The presence of microaneurysms (MAs) is usually an early sign of diabetic retinopathy and their automatic detection from color retinal images is of clinical interest. In this paper, we present a new approach for automatic MA detection from digital color fundus images. We formulate MA detection as a problem of target detection from clutter, where the probability of occurrence of target is considerably smaller compared to the clutter.[29]

Retinal image analysis is one of the profound areas of research. Computational solutions are sought in this context. Data Mining techniques have been extensively adopted for this purpose. In this work, Diabetic Retinopathy, a primary cause of blindness is dealt with. A two-level classification is adopted to classify Diabetic Retinopathy. In this classification, first level classification is performed through ensemble of Best First Trees.[30]

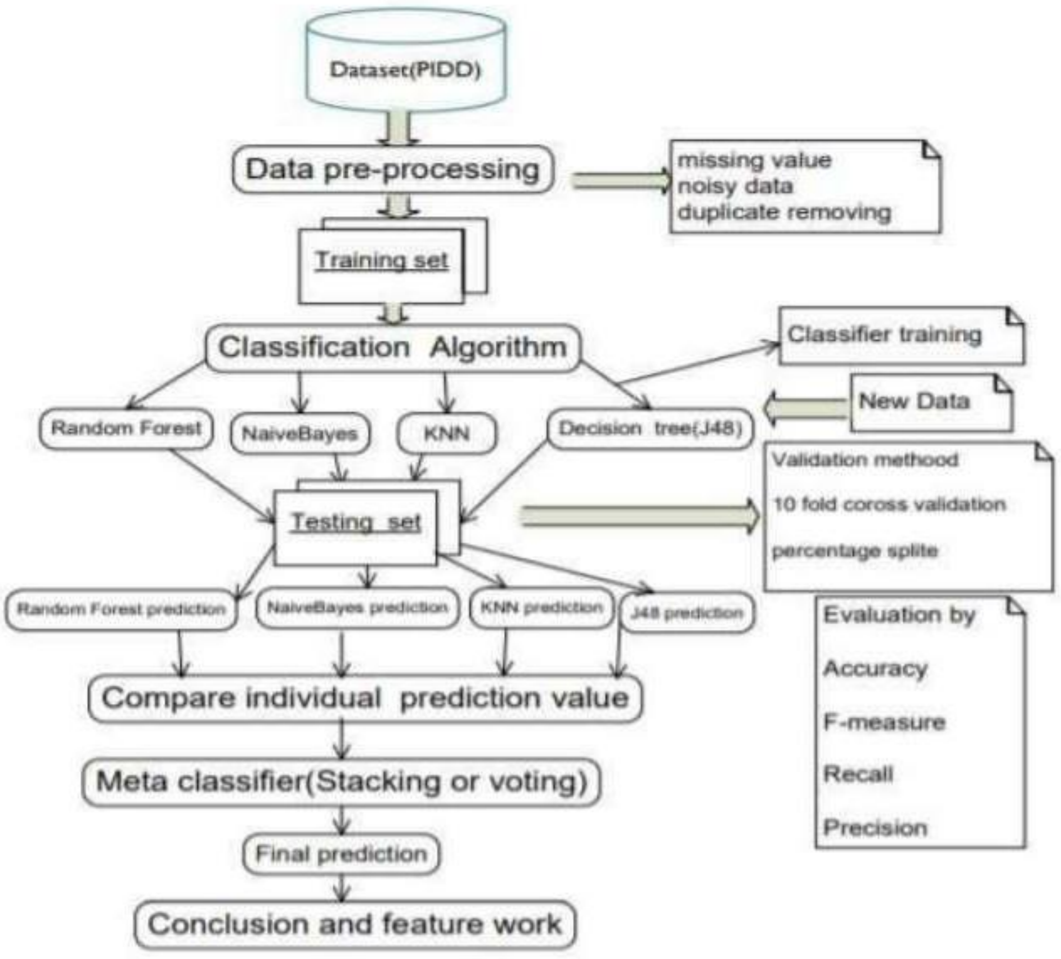
Field Name	Order	Type	Description
preg	1	Number	Number of times pregnant
plas	2	Number	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Pres	3	Number	Diastolic blood pressure (mm Hg)
Skin	4	Number	Triceps skin fold thickness (mm)
Insu	5	Number	2-Hour serum insulin (mu U/ml)
Mass	6	Number	Body mass index (weight in kg/(height in m)^2)
Pedi	7	Number	Diabetes pedigree function
Age	8	Number	Age (years)
class	9	string	Class variable (0 or 1)

Proposed system:

Ensemble modelling is a powerful way to improve the performance of your model. It usually pays off to apply ensemble learning over and above various models you might be building. Time and again, people have used ensemble models in competitions like Kaggle and benefited from it. Ensemble learning is a broad topic and is only confined by your own imagination. For the purpose of this article, I will cover the basic concepts and ideas of ensemble modeling. This should be enough for you to start building ensembles at your own end. As usual, we have tried to keep things as simple as possible. Let’s quickly start with an example to understand the basics of Ensemble learning. This example will bring out, how we use ensemble model every day without realizing that we are using ensemble modelling .

Ensemble is the art of combining diverse set of learners (individual models) together to improvise on the stability and predictive power of the model. In the above example, the way we combine all the predictions together will be termed as Ensemble Learning. In this article, we will talk about a few ensemble techniques widely used in the industry. Before we get into techniques, let’s first understand how do we actually get different set of learners. Models can be different from each other Diabetes mellitus (DM) or simply diabetes, is a group of metabolic diseases

Detailed Design:



Comparative study:

Classifier/Performance measure	Specificity	Sensitivity	Recall	F-Measure
KNN	0.51	0.88	0.77	0.82
Logistic Regression	0.54	0.87	0.78	0.82
Decision Tree	0.63	0.77	0.79	0.78
Naïve Bayes	0.52	0.85	0.77	0.81
Linear SVM	0.52	0.81	0.76	0.78
RBF SVM	0.52	0.88	0.77	0.82
Gaussian Process	0.52	0.86	0.77	0.81
Ada Boost	0.57	0.82	0.78	0.80
Random Forest	0.63	0.83	0.81	0.82
Voting Classifier(30% test data)	0.60	0.92	0.81	0.86

Algorithm used :

DATA CLUSTERING DATASET LINK : -

Data Clustering is a technique of data division into separate parts or sections that partitions the data into several groups based on their similarity and same data. Basically, we group the data through a statistical operation and information that withholds on that operation. These smaller groups that are combined from the bigger data are known as clusters

step 1 : find the decision tree of the dataset .

first load the data set into the r studio and in model set to tree and than click on the execute then you will find draw option on the screen.if you click on the draw option you can see the decision tree diagram

step 2 : select the leaf node.

In the decision tree you see the nodes. For each node you can see the values of the data instruction with that instruction edit the main data set and copy that data and mark it as leaf node.

step 3 : filter the leaf node according to the decision tree.

For leaf node you should follow that step 2 process

step 4 : find the number of nodes in the child node.

While calculating the error in the edited data set. We should check whether the tree contains child node. We should also add those instructions to while doing step 2.

step 5 : find the proposition of the child node.

proposition = no. of nodes/total no. of nodes

step 6 : find the error percentage of the child node.

Load the leaf node into the r studio and click on evaluate and make it full and then execute the data set.

step 7 : multiply the error percentage and proposition'

in step 5 with the formula we can find the proposition value and error percentage can find following the sep 6

step 8 : apply the same process for all leaf nodes

continue the process from step 1 to step 7 to find final error percentage.

R-Platform

We use R software to predict the result using all the algorithms. All the results will be displayed in graphs, decision trees using the R software. Why R tool?

1. It is one step process to calculate any Statistics,

2. It is one step process to implement any Data Mining/ Machine Learning Algorithms

3. It is one step process to Visualize the data.

4. R can read data from any Data Source.

5. R Visuals and outputs can be integrated with any environment.

6. Rich R packages. Any technique you imagine, there will be an R package to implement that here.

7. It gives Easy framework for the Data Scientists from different background to experiment, analyse and visualize data.

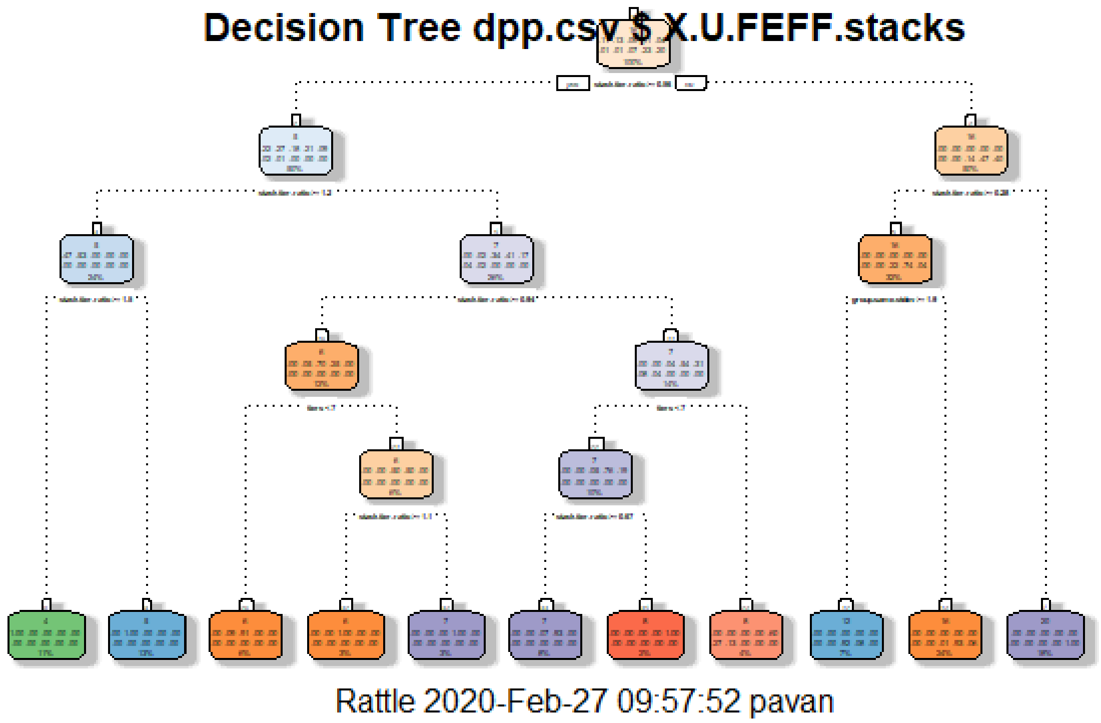
DECISION TREE

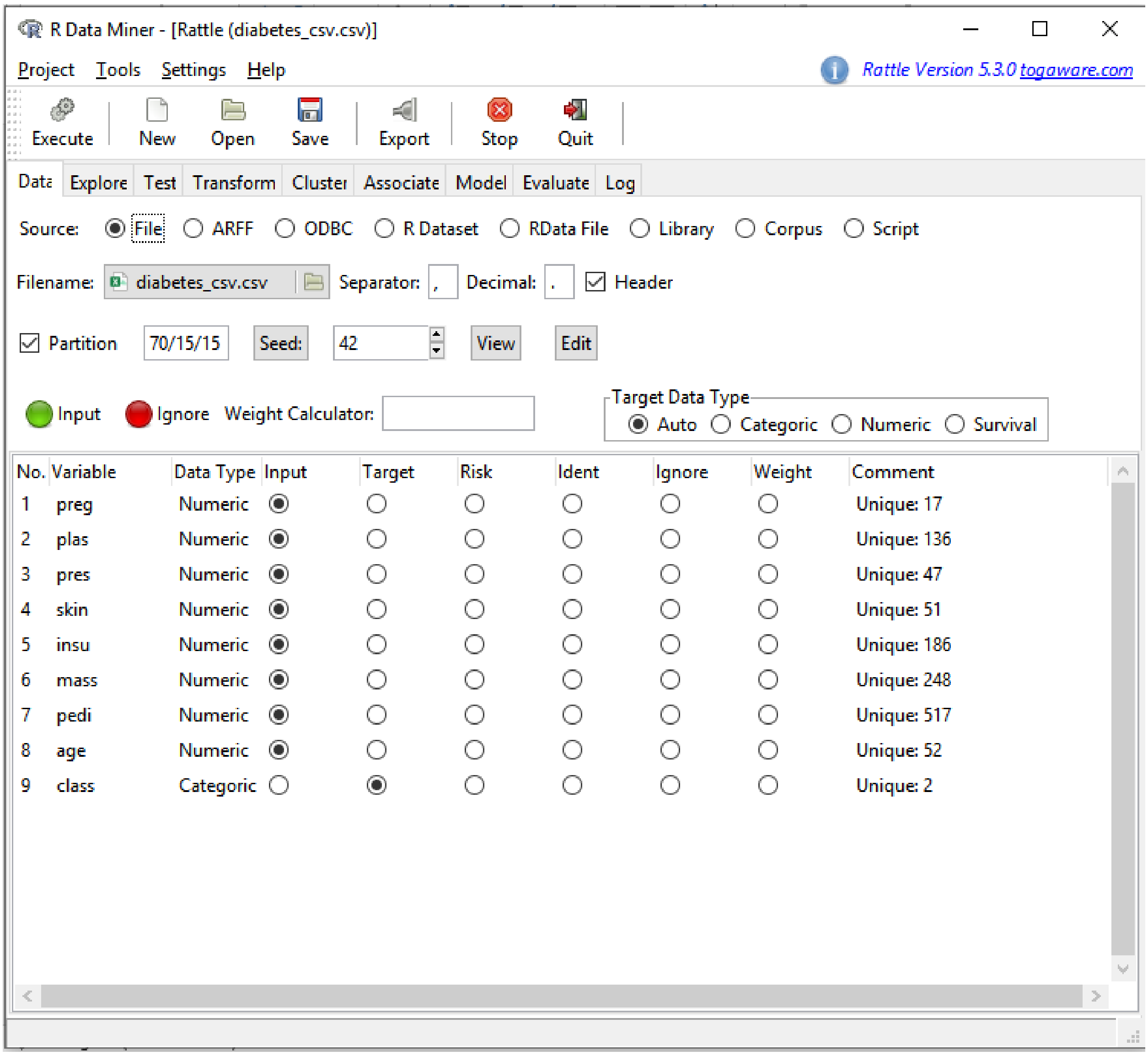
1) it contains 10 attributes

2) 1097 instances presented in the decision tree

3) a total of 12 nodes are present.

4) this helps us to calculate the leaf nodes from the data set and while calculating the leaf nodes we should also keep in mind that other nodes are connected.


Decision Tree dpp.csv $ X.U.FEFF.stacks
yes
no
Rattle 2020-Feb-27 09:57:52 pavan


R Data Miner - [Rattle (diabetes_csv.csv)]
Project Tools Settings Help
Rattle Version 5.3.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ● File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

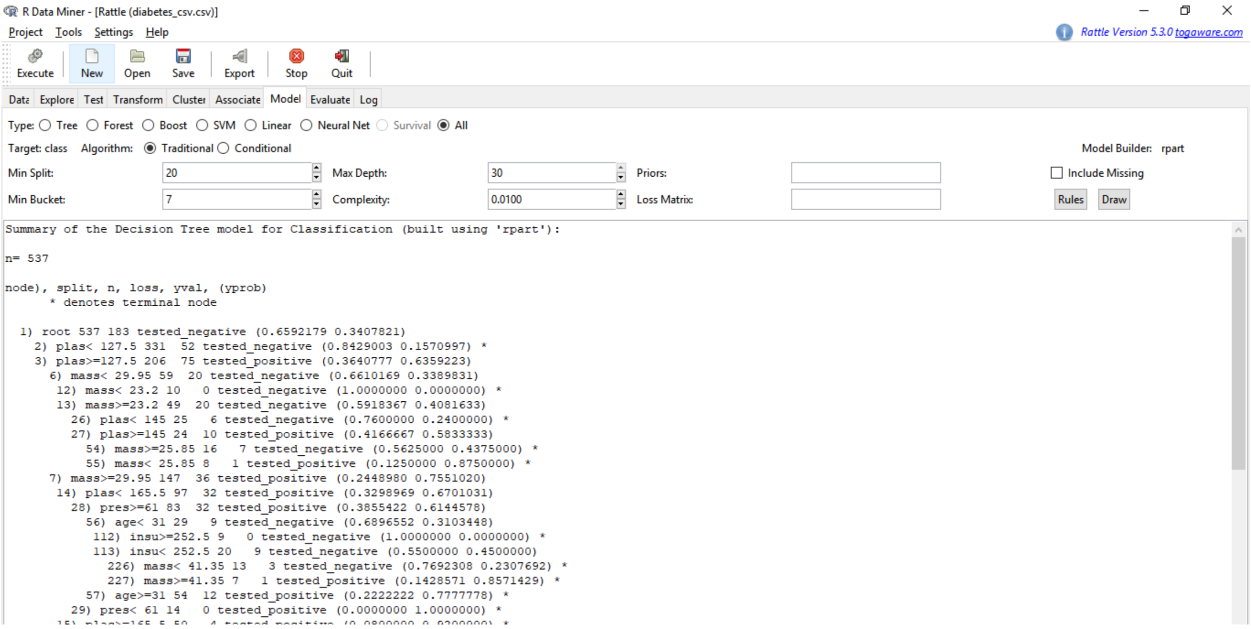
Filename: diabetes_csv.csv Separator: , Decimal: . ☑ Header

☑ Partition 70/15/15 Seed: 42 View Edit

● Input ● Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	preg	Numeric	●	○	○	○	○	○	Unique: 17
2	plas	Numeric	●	○	○	○	○	○	Unique: 136
3	pres	Numeric	●	○	○	○	○	○	Unique: 47
4	skin	Numeric	●	○	○	○	○	○	Unique: 51
5	insu	Numeric	●	○	○	○	○	○	Unique: 186
6	mass	Numeric	●	○	○	○	○	○	Unique: 248
7	pedi	Numeric	●	○	○	○	○	○	Unique: 517
8	age	Numeric	●	○	○	○	○	○	Unique: 52
9	class	Categoric	○	●	○	○	○	○	Unique: 2


R Data Miner - [Rattle (diabetes_csv.csv)]
Project Tools Settings Help
Rattle Version 5.3.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rules Draw

Type: ○ Tree ○ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ● All

Target: class Algorithm: ● Traditional ○ Conditional

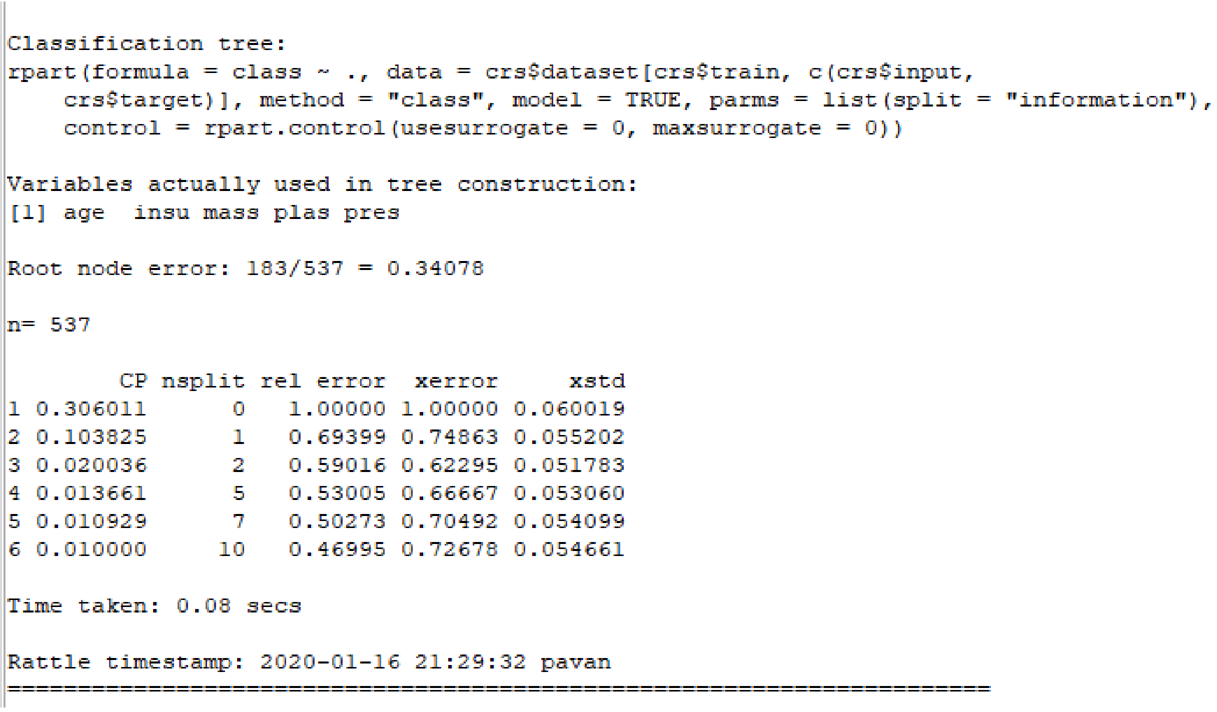
Model Builder: rpart

Min Split: 20 Max Depth: 30 Priors:

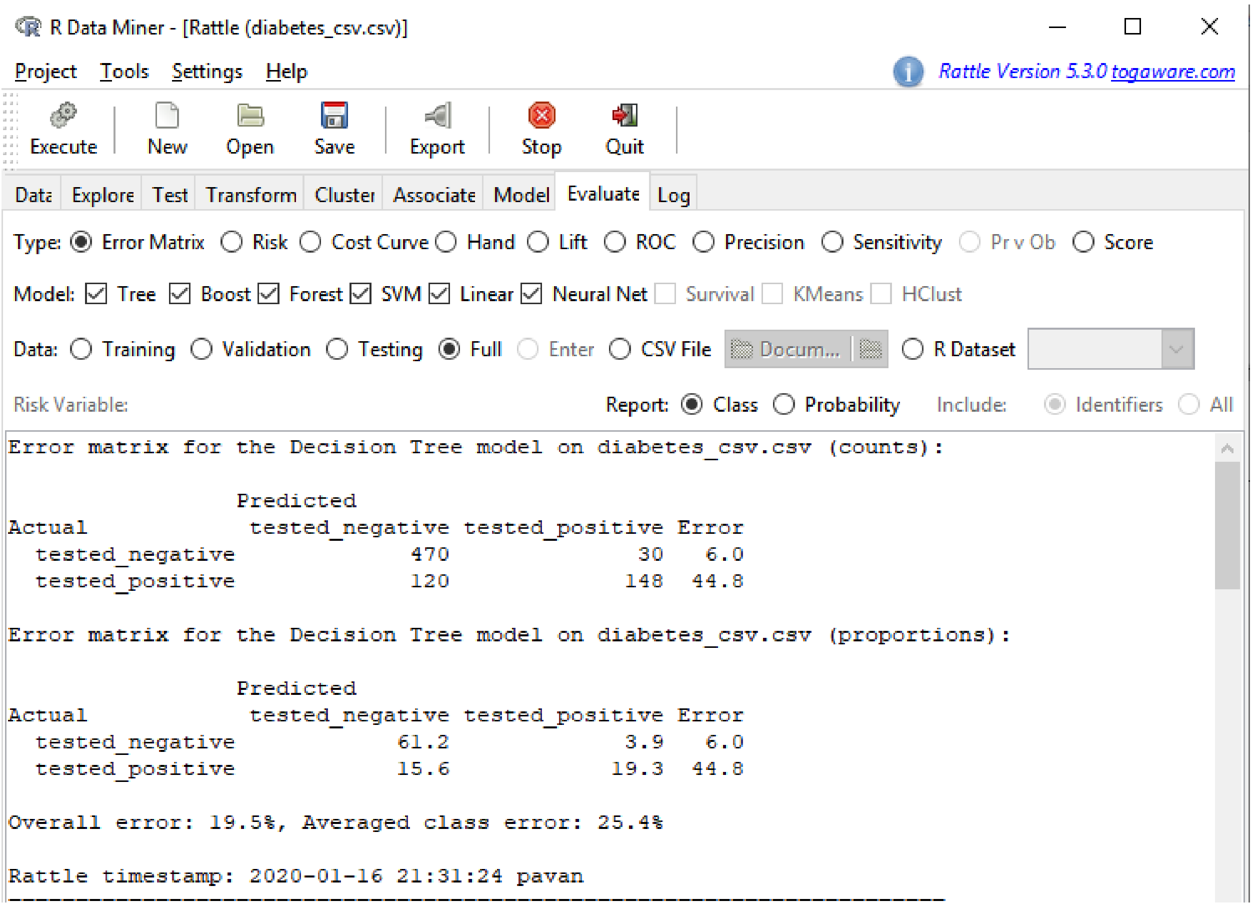
Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

□ Include Missing

```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 537

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 537 183 tested_negative (0.6592179 0.3407821)
    2) plas< 127.5 331  52 tested_negative (0.8429003 0.1570997) *
    3) plas>=127.5 206  75 tested_positive (0.3640777 0.6359223)
      6) mass< 29.95 59  20 tested_negative (0.6610169 0.3389831)
       12) mass< 23.2 10   0 tested_negative (1.0000000 0.0000000) *
       13) mass>=23.2 49  20 tested_negative (0.5918367 0.4081633)
         26) plas< 145 25   6 tested_negative (0.7600000 0.2400000) *
         27) plas>=145 24  10 tested_positive (0.4166667 0.5833333)
           54) mass>=25.85 16   7 tested_negative (0.5625000 0.4375000) *
           55) mass< 25.85 8   1 tested_positive (0.1250000 0.8750000) *
      7) mass>=29.95 147  36 tested_positive (0.2448980 0.7551020)
       14) plas< 165.5 97  32 tested_positive (0.3298969 0.6701031)
         28) pres>=61 83  32 tested_positive (0.3855422 0.6144578)
           56) age< 31 29   9 tested_negative (0.6896552 0.3103448)
            112) insu>=252.5 9   0 tested_negative (1.0000000 0.0000000) *
            113) insu< 252.5 20   9 tested_negative (0.5500000 0.4500000)
              226) mass< 41.35 13   3 tested_negative (0.7692308 0.2307692) *
              227) mass>=41.35 7   1 tested_positive (0.1428571 0.8571429) *
           57) age>=31 54  12 tested_positive (0.2222222 0.7777778) *
         29) pres< 61 14   0 tested_positive (0.0000000 1.0000000) *
       15) plas>=165.5 50   4 tested_positive (0.0800000 0.9200000) *
```



```
Classification tree:
rpart(formula = class ~ ., data = crs$dataset[crs$train, c(crs$input, 
    crs$target)], method = "class", model = TRUE, parms = list(split = "information"), 
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] age  insu mass plas pres

Root node error: 183/537 = 0.34078

n= 537 

        CP nsplit rel error  xerror     xstd
1 0.306011      0   1.00000 1.00000 0.060019
2 0.103825      1   0.69399 0.74863 0.055202
3 0.020036      2   0.59016 0.62295 0.051783
4 0.013661      5   0.53005 0.66667 0.053060
5 0.010929      7   0.50273 0.70492 0.054099
6 0.010000     10   0.46995 0.72678 0.054661

Time taken: 0.08 secs

Rattle timestamp: 2020-01-16 21:29:32 pavan
======================================================================
```


R Data Miner - [Rattle (diabetes_csv.csv)]
Project Tools Settings Help
Rattle Version 5.3.0 togaware.com
Execute
New
Open
Save
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

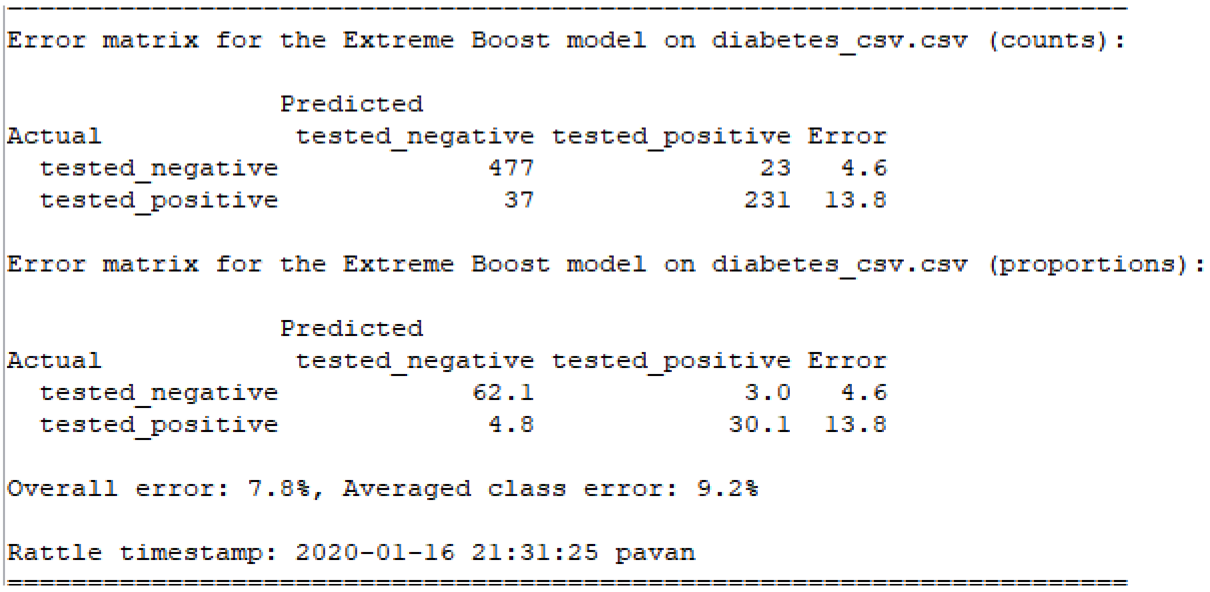
Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

Model: ☑ Tree ☑ Boost ☑ Forest ☑ SVM ☑ Linear ☑ Neural Net □ Survival □ KMeans □ HClust

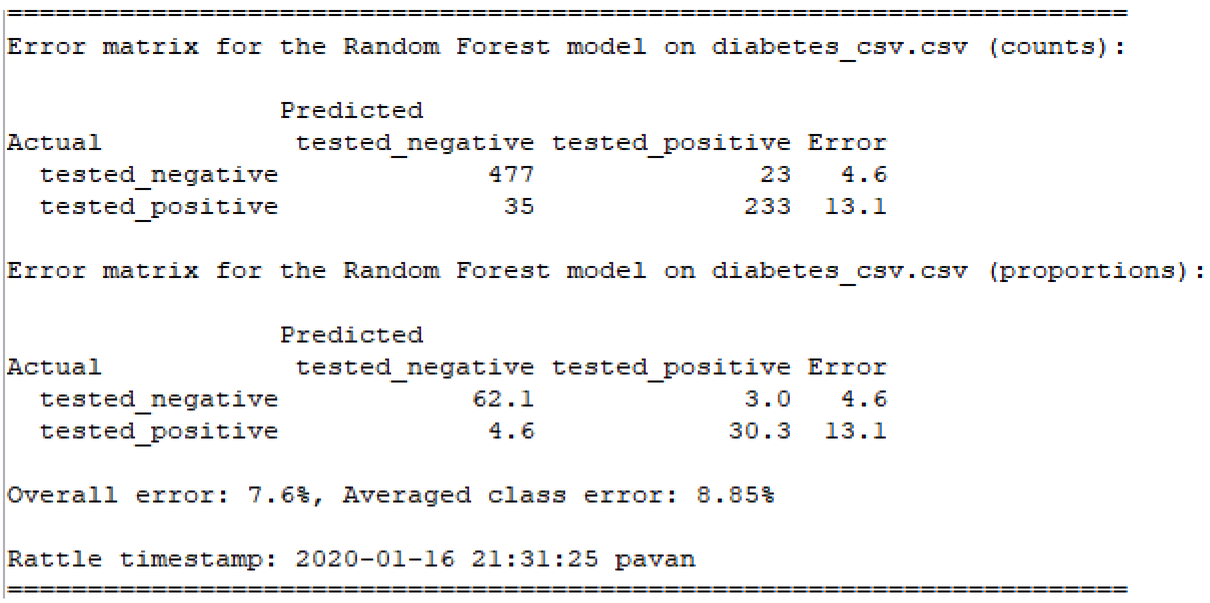
Data: ○ Training ○ Validation ○ Testing ● Full ○ Enter ○ CSV File Docum... ○ R Dataset

Risk Variable: Report: ● Class ○ Probability Include: ● Identifiers ○ All

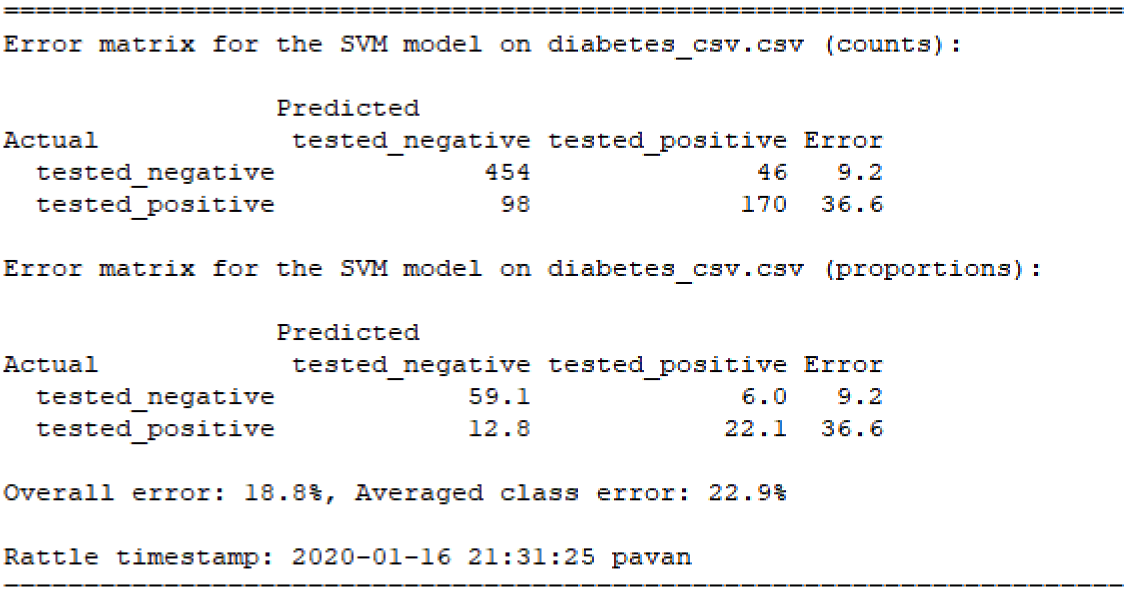
```
Error matrix for the Decision Tree model on diabetes_csv.csv (counts):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative             470              30   6.0
  tested_positive             120             148  44.8

Error matrix for the Decision Tree model on diabetes_csv.csv (proportions):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative            61.2             3.9   6.0
  tested_positive            15.6            19.3  44.8

Overall error: 19.5%, Averaged class error: 25.4%

Rattle timestamp: 2020-01-16 21:31:24 pavan
----------------------------------------------------------------------
```



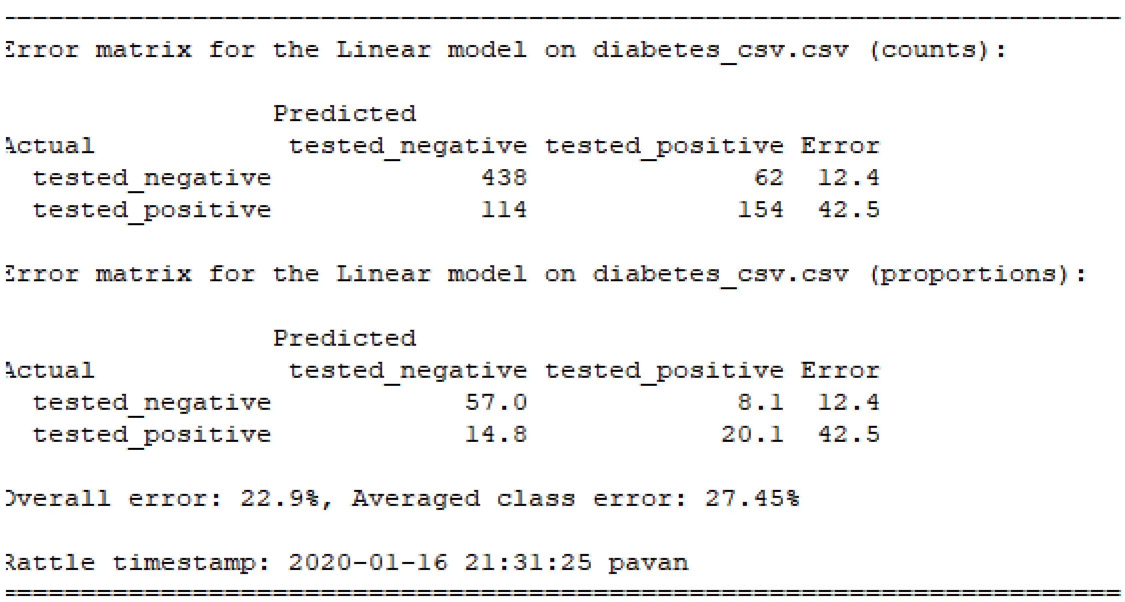
```
----------------------------------------------------------------------
Error matrix for the Extreme Boost model on diabetes_csv.csv (counts):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative             477              23   4.6
  tested_positive              37             231  13.8

Error matrix for the Extreme Boost model on diabetes_csv.csv (proportions):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative            62.1             3.0   4.6
  tested_positive             4.8            30.1  13.8

Overall error: 7.8%, Averaged class error: 9.2%

Rattle timestamp: 2020-01-16 21:31:25 pavan
======================================================================
```



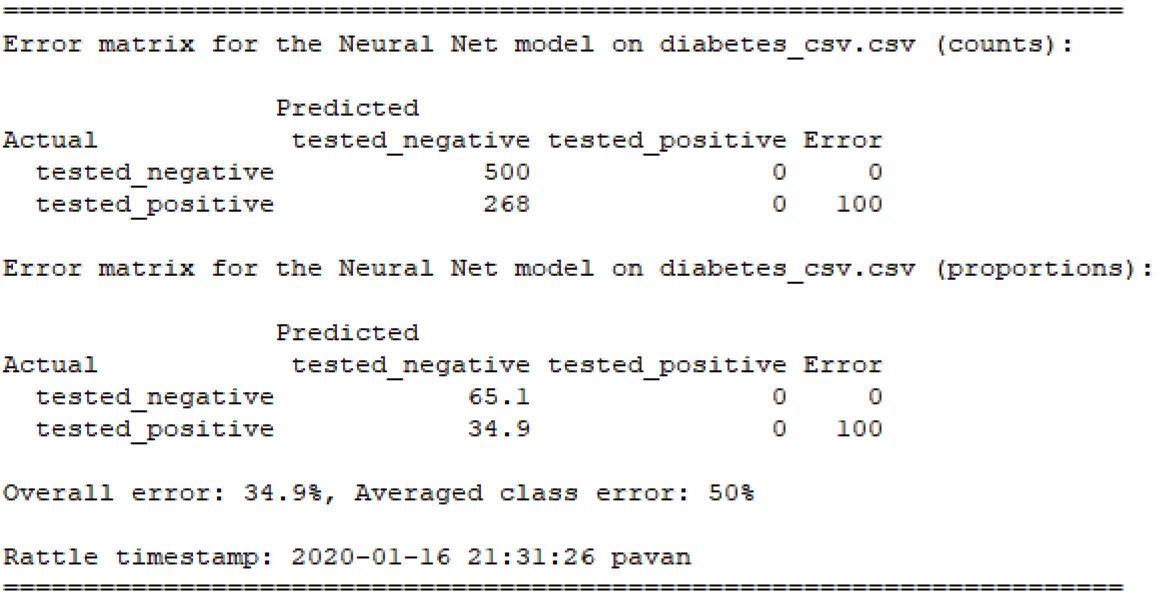
```
======================================================================
Error matrix for the Random Forest model on diabetes_csv.csv (counts):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative             477              23   4.6
  tested_positive              35             233  13.1

Error matrix for the Random Forest model on diabetes_csv.csv (proportions):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative            62.1             3.0   4.6
  tested_positive             4.6            30.3  13.1

Overall error: 7.6%, Averaged class error: 8.85%

Rattle timestamp: 2020-01-16 21:31:25 pavan
======================================================================
```



```
======================================================================
Error matrix for the SVM model on diabetes_csv.csv (counts):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative             454              46   9.2
  tested_positive              98             170  36.6

Error matrix for the SVM model on diabetes_csv.csv (proportions):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative            59.1             6.0   9.2
  tested_positive            12.8            22.1  36.6

Overall error: 18.8%, Averaged class error: 22.9%

Rattle timestamp: 2020-01-16 21:31:25 pavan
----------------------------------------------------------------------
```



```
======================================================================
Error matrix for the Linear model on diabetes_csv.csv (counts):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative             438              62  12.4
  tested_positive             114             154  42.5

Error matrix for the Linear model on diabetes_csv.csv (proportions):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative            57.0             8.1  12.4
  tested_positive            14.8            20.1  42.5

Overall error: 22.9%, Averaged class error: 27.45%

Rattle timestamp: 2020-01-16 21:31:25 pavan
======================================================================
```



```
======================================================================
Error matrix for the Neural Net model on diabetes_csv.csv (counts):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative             500               0     0
  tested_positive             268               0   100

Error matrix for the Neural Net model on diabetes_csv.csv (proportions):

                 Predicted
Actual            tested_negative tested_positive Error
  tested_negative            65.1               0     0
  tested_positive            34.9               0   100

Overall error: 34.9%, Averaged class error: 50%

Rattle timestamp: 2020-01-16 21:31:26 pavan
======================================================================
```

Result:

Split.No	Decision Tree	Random Forest	SVM	Linear	Neural Network	Best Model
D1	7.1	4.2	10.1	10.4	-89.2	Random Forest
D2	0.8	0.8	-	1.7	-99.2	Decision tree, Random Forest
D3	1.9	0	-	1.9	-98.1	Random Forest
D4	21.7	14.9	25.7	27.5	24.8	Random Forest
D5	29.4	21.2	30.9	31.2	28.5	Random Forest
D6	6.4	4	12.2	15.7	-62.8	Random Forest
D7	6.5	4.3	6.5	4.4	-93.5	Random Forest
D8	14.7	3	10.5	10.5	-79.1	Random Forest
D9	18.5	3.7	18.5	21	-21	Random Forest
D10	15.7	5	7.4	6.6	-15.7	Random Forest
D11	13.3	0	-	-86.7	0	Random Forest
D12	10.5	2.7	8.7	13.7	-13.5	Random Forest
Average Error	12.208	5.31	14.5	19.275	52.11	Random Forest

ERROR CALCULATIONS : -

Overall Error rate of Best Performing Model:

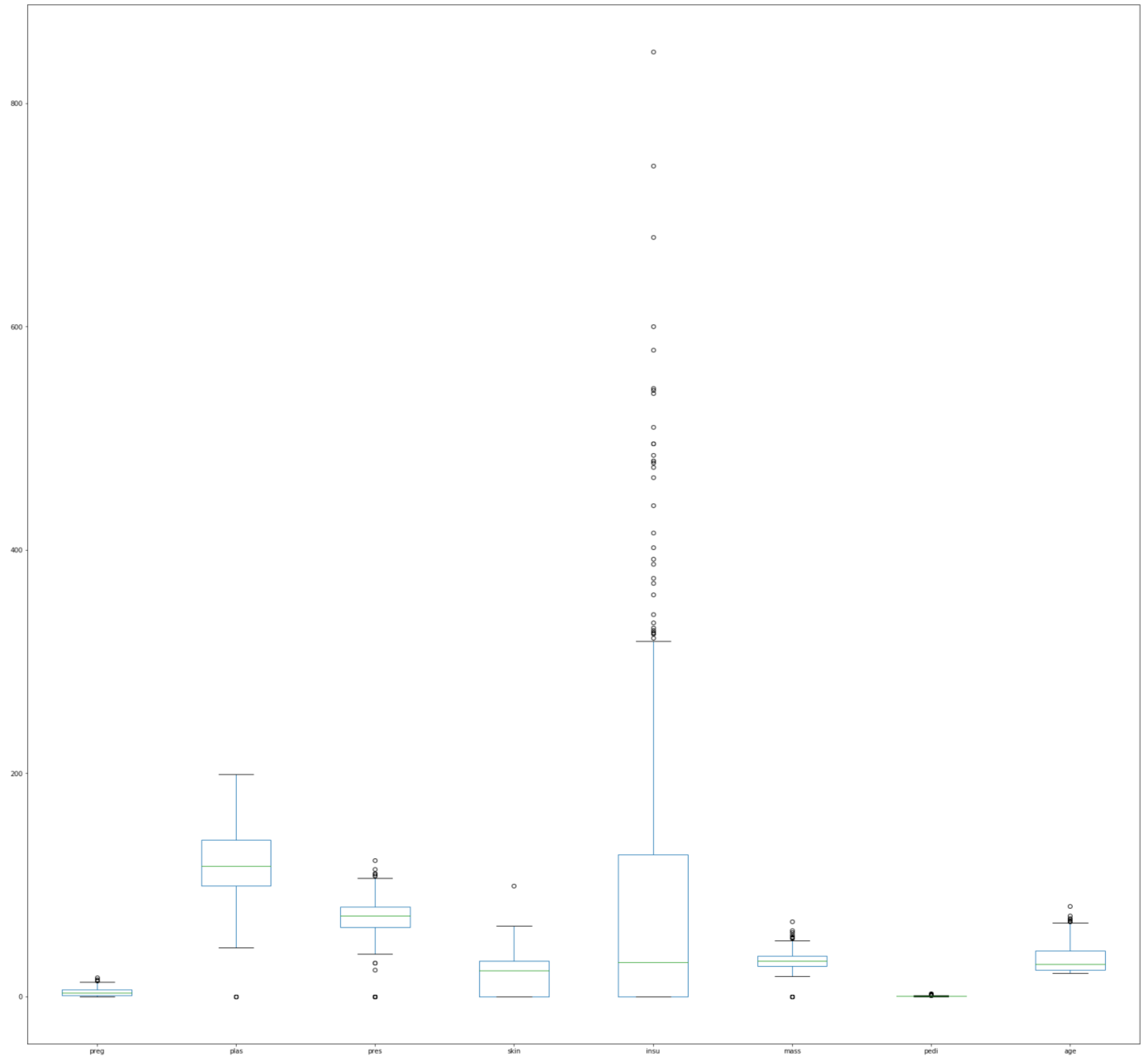
Best error rate of data set is: **5.31**

Best Model: **Random Forest Model**

Malpolt graph:

```
import matplotlib.pyplot as plt
d.plot(kind ='Box' , figsize = (30,30))
plt.show()
```

Graph:



References:

[1] Autonomic neuropathy in diabetes: early detection and the role in development of microvascular complications. A. Bondar, V. V. Klimontov and
E. A. Korolyova

[2] Detection of Hypoglycemic Episodes in Children with Type 1 Diabetes using an Optimal Bayesian Neural Network Algorithm. Hung T. Nguyen, Senior Member IEEE, NejhdehGhevondia n, Son T. Nguyen, Timothy W. Jones

[3] Detection of diabetes using genetic programming. Muhammad Waqar Aslam and Asoke Kumar Nandi The University of Liverpool.

[4] A Classifier Based Approach for Early Detection of Diabetes Mellitus. Sumangali K.

[5] A computation analysis to predict diabetes based on data mining. Girdhar Gopal Ladha PhD Scholar, Department of Computer Science RKDF University, Bhopal

[6] Detection and Prediction of Diabetes Mellitus Using Back- Propagation Neural Network.Miss. Sneha Joshi Prof. MeghaBorse

[7] Early Detection of Diabetic Retinopathy from Digital Retinal Fundus Images.

Deepthi K Prasad, Vibha L, Venugopal K R.

[8] Application of Data Mining Techniques to Predict the Length of Stay of Hospitalized Patients with Diabetes. Ayman Alahmar, Emad A.Mohammed, Rachid Benlamri.

[9] Early detection of diabetes mellitus disease Using Artificial Neural Network Bakpropagation with adaptive learning rate and particle swarn optimization.FikerAofa priyo sidisongaos utikno

[10] A review and analysis on data mining methods to predict diabetes .
GirdharGopal Ladha Ravi Kumar Singh Pippal

[11] computer-assisted non-invasive diabetes mellitus detection system via facial key block analysis. Tingshu,bobzhang,yu an-yantang

[12] Prediction of Diabetes Using Bayesian Network. Mukesh kumari , Dr.

Rajan Vohra,Anshul arora.

[13] Design and develop an algorithm for a diabetic detection using ECG signal. ReenaMusale, A. N.Paithane

[14] Prediction of Diabetes using EnsembleTechniques. Prema N S, Varshith V, Yogeswar J

[15] Prediction of diabetes using artificial network approach.suyash Srivastav,Lokesh sharma,Dr.Ajai kumar, and Dr.Hemant Derbari.

[16] Disease Influence Measure Based Diabetic Prediction with Medical Data Set Using Data Mining. B.V. Baiju and Dr.D. John Aravindhar

[17] Prediction of heart and kidney risks in Diabetic Prone Population using Fuzzy Classification. S.Ananthi nad V.Bhuvaneswari

[18] Implementation of Diabetic Retinopathy Prediction System using Data Mining. S. Patil and Prof. Kalpana Malpe

[19] An efficient prediction of diabetic from retinopathy using machine learning and signal processing approach . Kalyan Kumar Mohanty, Prabhat Kumar Barik, Ram Chandra Barik, and Ram Chandra Barik,

[20] Extraction of Blood Vascular Network for Development of an Automated Diabetic Retinopathy Screening System. S. Jerald Jeba Kumar and M.Madheswaran

[21] Improvement in Prediction Rate and Accuracy of Diabetic Diagnosis System Using Fuzzy Logic Hybrid Combination. Poonam Undre, Harjeet Kaur and Prakash Patil.

[22] Feasibility of Diagnosing Both Severity and Features of Diabetic Retinopathy in Fundus Photography. Juan Wang, Yujing Bai, and Bin Xia.

[23] Predicting Serious Diabetic Complications using Hidden Pattern Detection. Saeed Farzi, Sahar Kianian, and Ilnaz Rastkhadive.

[24] Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. MD. Kamrul Hasan , MD. Ashraful Alam , Dola Das , Eklas Hossain , and Mahmudul Hasan

[25] Automatic Diabetic Retinopathy Detection and Classification System. Z. A. Omar, M. Hanafi, S. Mashohor, N. F. M. Mahfudz and M.

Muna’im.

[26] Predicting blood glucose levels in diabetics using feature extraction and Artificial Neural Networks. Khaled Eskaf, Prof. Dr. Osama Badawi and Prof. Dr. Tim Ritchings.

[27] Diabetic Retinopathy Detection by Extracting Area and Number of Microaneurysm from Colour Fundus Image. Shailesh Kumar and Basant Kumar.

[28] Design and Implementation of a Unique Blood-vessel Detection Algorithm towards Early Diagnosis of Diabetic Retinopathy. Sumeet Dua, Naveen Kandiraju and W. Thompson.

[29] A Successive Clutter-Rejection-Based Approach for Early Detection of Diabetic Retinopathy. Keerthi Ram, Gopal Datt Joshi and Jayanthi Sivaswamy.

[30] Automatic Diabetic Retinopathy Detection through Ensemble Classification Techniques. Dr. R. GeethaRamani, Jeslin Shanthamalar J and Lakshmi B.